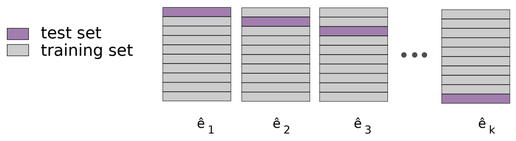
Look before you leap:
Some insights into learner evaluation with cross-validation

Gitte Vanwinckelen and Hendrik Blockeel
KU Leuven, Belgium

Context and motivation

Task: Evaluation of supervised learning algorithms with cross-validation

k-fold cross-validation: $\hat{e}_{cv} = \frac{1}{k}\sum \hat{e}_i$



variants

- leave-one-out cross-validation: k=number of instances in D
- repeated cross-validation: $\hat{e}_{rcv} = \frac{1}{r}\sum \hat{e}_{cv}$

This task definition is, however, still somewhat vague:

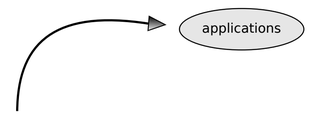
- Which error are we estimating?
- How reliable is our cross-validation error estimate?

Our **goal** is to answer these questions and give insight into the correct use of cross-validation, and the correct interpretation of the resulting error estimate

Error measures

The error of a classifier m is the probabiity of making an incorrect prediction for an instance drawn randomly from the population P: $e(m) = Pr_{(x,y)\sim p}[m(x) \neq y]$

Two types of errors can be distinguished for a **learner** L



1. Conditional error $e_c = e(L,T)$

   Given a dataset D from population P, and a set of learners, which learner learns from D the most accurate model on P?

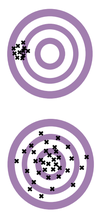
2. Unconditional error $e_u = \underset{\{S:\,|S|=n\}}{E}[e(L,T)]$

   Given a population P, and a set of learners, which learner is expected to yield the most accurate model on P, when given a random sample of a particular size from P?

Quality of the cross-validation estimator

The quality of an estimator ê is expressed by its Mean Squared Error: $MSE(\hat{e},e) = \underset{S,F}{E}[(e-\hat{e})^2]$

$MSE(\hat{e},e) = Variance(\hat{e}) + Bias^2(\hat{e},e)$



✓ $Bias(\hat{e},e) = \underset{S,F}{E}[\hat{e}-e]$

✓ $Variance(\hat{e}) = \underset{S,F}{E}[(\hat{e}-\underset{S,F}{E}[\hat{e}])^2]$

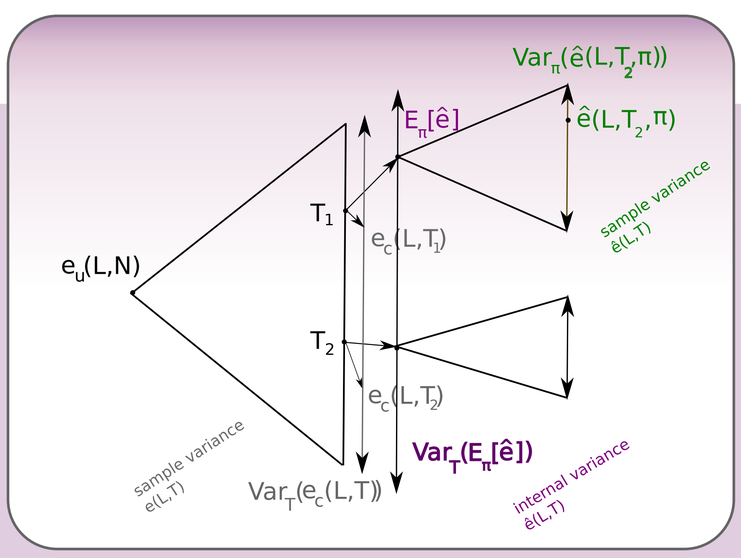
Estimating the variance of the cross-validation estimator is a difficult problem

If each evaluation on a test fold would be an independent evaluation: $Var(\hat{e}) = \frac{Var(\hat{e}_i)}{k}$

However, the actual variance of ê is higher because of dependencies between [4]:

✓ errors on different test folds, caused by partially overlapping training folds

✓ errors on the same test fold, caused by evaluating the same model



✓ The **variance of the cross-validation** estimator depends on two random variables: The sample and the fold partitioning. Therefore [2,3]:

$$Var_{S,F}(\hat{e}) = Var_S(E_F[\hat{e}|S]) + E_S[Var_F(\hat{e}|S)] \quad \textit{(law of total variance)}$$

With one sample we can only estimate $Var_F(\hat{e}|S)$.
We do not know how much the error estimate varies over all possible samples [2,5]

$Var_F(\hat{e}|S)$ cannot replace the sample variance: It reduces to 0 when averaging over all fold partitionings

$Var_F(\hat{e}|S)$ is a property of the estimation method, not the learning problem

✓ When varying both sample and partitioning, cross-validation is unbiased for the unconditional error (ignoring the bias because of smaller training sets)

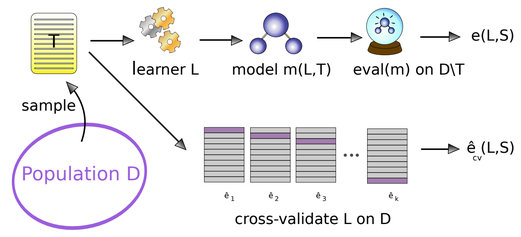
However, when fixing sample T, while varying its partitioning, there is **no guarantee that $E_F[\hat{e}|S] = e(L,T)$**

Stated otherwise, the repeated cross-validation estimate does not always converge to the conditional or the unconditional error, when averaging over an increasing number of partitionings

Experimental setup

Procedure

Given: A large dataset D (population) and a learner L
For i from 1 to M do:
  Sample a small dataset T with n instances from D
  1. Compute e(L,T) by learning a model on T and evaluating it on D \ T
  2. Compute $\hat{e}_{cv}(L,T)$ by using cross-validation

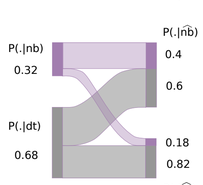


Comparing learners with cross-validation

How often is the best learner, when evaluated with cross-validation, also the learner with the best conditional error?

Evaluation of naive Bayes and a decision tree on 100 samples from the adult UCI [1] dataset with twofold cross-validation:

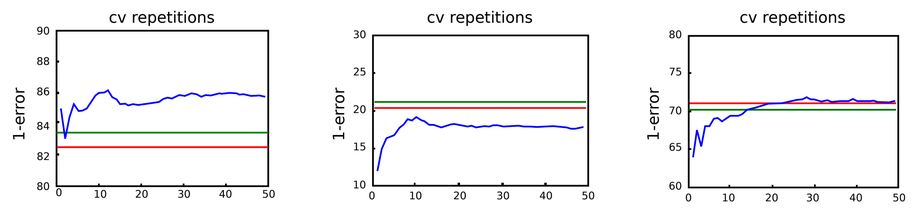
	nb	dt
$\widehat{nb}$	25	37
$\widehat{dt}$	7	31



✓ When the decision tree outperforms naive Bayes, cross-validation only selects the decision tree with P = 0.54

✓ When cross-validation selects naive Bayes, the probability is only 0.4 that naive Bayes is actually better
This is because of the many samples on which the decision tree wins, where yet naive Bayes is selected as the winner

How well does repeated cross-validation estimate the error?



✓ In many cases, ê(L,T) does not converge to the conditional, nor to the unconditional error, when increasing the number of repetitions

✓ Yet, 10 to 20 repetitions often increases the accuracy of the error estimate for either parameter. This is because of the reduced internal variance when averaging over multiple cv repetitions

Conclusions

✓ It should always clearly be stated whether one is estimating the error of the learner or the model

✓ Repeated cross-validation leads to more accurate estimates of the error of both the learner and the model

✓ Repeated cross-validation does not result in more accurate statistical inference

KU LEUVEN

DTAI
Declaratieve Talen en Artificiële Intelligentie

Future work

Our experiments indicate that in a few cases the repeated cross-validation estimate converges to the conditional or the unconditional error. Can we determine the properties of the learning problems for which this is the case?

References

1. Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. Hanczar B, Dougherty ER (2010) On the comparison of classifiers for microarray data. Current Bioinformatics 5:29–39
3. J. D. Rodriguez, A. Perez, and J. A. Lozano (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. 32, no. 3, pp. 569–575.
4. Bengio Y, Grandvalet Y (2004) No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research 5:1089–1105
5. Isaksson A, Wallman M, G¨ransson H, Gustafsson MG (2008) Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters 29(14):1960–1965